DOCUMENT RESUME

ED 478 314 SE 068 145

| | |
|---|---|
| AUTHOR | Lesh, Richard; Clarke, David |
| TITLE | Formulating Operational Definitions of Desired Outcomes of Instruction in Mathematics and Science Education. |
| PUB DATE | 2000-00-00 |
| NOTE | 38p.; In: Kelly, Anthony E. and Richard Lesh, Handbook of Research Design in Mathematics and Science Education, Mahwah, NJ, Lawrence Erlbaum, 1999. p113-49. |
| AVAILABLE FROM | Lawrence Erlbaum Associates, Inc., 10 Industrial Avenue, Mahwah, NJ 07430. Web site: http://www.erlbaum.com. |
| PUB TYPE | Books (010) -- Reports - Descriptive (141) |
| EDRS PRICE | EDRS Price MF01 Plus Postage. PC Not Available from EDRS |
| DESCRIPTORS | Alternative Assessment; *Curriculum Development; *Educational Change; Educational Objectives; Elementary Secondary Education; *Evaluation; *Science Education; *Standardized Tests |

ABSTRACT

This chapter discusses alternatives beyond traditional standardized tests or curriculum standards consisting of naive checklists of low-level facts and skills to operationally define educational goals in ways that provide direction for both instruction and accountability. Suggestions are drawn from physical and social sciences as well as from common experiences in business, engineering, and other areas where science is applied. (KHR)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 478 314

# 6 Formulating Operational Definitions of Desired Outcomes of Instruction in Mathematics and Science Education

**Richard Lesh**
*Purdue University*

**David Clarke**
*University of Melbourne*

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GHANTED BY

S. Barker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The central question this chapter is intended to answer is: "What other options are available, beyond traditional standardized tests, or curriculum standards consisting of naïve checklists of low-level facts and skills, to operationally define educational goals in ways that provide direction for both instruction and accountability?" Suggestions for alternatives will be drawn from both the physical and social sciences, as well as from common experiences in business, engineering, and other areas where science is applied.

## LIMITATIONS OF STANDARDIZED TESTS TO OPERATIONALLY DEFINE GOALS OF INSTRUCTION

Productive assessment involves clarifying goals; and, the goal of this chapter is to describe several ways that mathematics and science educators can approach the task of developing measurable ways to specify the nature of their most important goals for instruction. To highlight some important issues that arise in such efforts, it is useful to begin by sorting out distinctions among a variety of assessment-related terms that educators often treat as if their meanings were identical—even though they are interpreted in ways that are quite distinct from one another in many other professions (ranging from business to engineering) or sciences (ranging from physical sciences to life sciences). By sorting out such distinctions, we believe

SE068145



BEST COPY AVAILABLE

that several significant needs, opportunities, and errors become apparent that should be taken into account when developing new perspectives about "operational definitions" of educational goals.

TABLE 6.1

| *Terms Commonly Emphasized in Traditional Testing* | *Terms Commonly Emphasized in Performance Assessment* |
|---|---|
| *Testing*<br>creating an ordeal (barrier, or filter) for accepting or rejecting (but not helping) | *Examining*<br>inspecting closely and generating high-fidelity descriptions |
| *Measuring*<br>partitioning (fragmenting) into indistinguishable pieces | *Documenting*<br>gathering tangible evidence which is credible to decision-makers |
| *Evaluating*<br>assigning a value without specifying conditions or purposes | *Assessing*<br>taking stock, orienting with respect to known landmarks and goals |

As Table 6.1 suggests, it is possible to inspect students' work closely without putting them through an ordeal; and, it also is possible to put students through an ordeal without closely inspecting their work. Similarly, it is possible to measure students without documenting their progress in forms that are useful to most decision makers; or, it is possible to document progress along a variety of dimensions without specifying a common unit of measure—and without collapsing all comparisons into a single linearly ordered dimension. Finally, it is possible to make value judgements (about a position or path) without getting oriented; and, it is possible to get oriented without making value judgements. For example, when travelers see an arrow on a map that says "you are here," they may be able to *assess* whether progress has been made toward some goal. Yet, judgements about whether a given location is good or bad may depend on a variety of factors. For example, if the travelers are hiking through the Grand Canyon, and if they are making judgements about how close they are to some specific goal, then these decisions may depend on factors related to time or effort rather than a geometer's notion of the straight-line distance between two points. Under some

circumstances, time may be more important than effort; or, under other circumstances, the opposite may be the case.

Similar points are apparent when a person uses a consumer guidebook to compare automobiles, restaurants, or vacation plans. It is neither necessary nor desirable for the guidebook to reduce comparisons to a single dimension, nor is it necessary to compare items using a single standard unit of measure. Is a Jeep Cherokee better than a Ford Taurus? Judgements about "better" and "worse" beg the question "Better for what?" — Clearly, answers to this question vary from person to person, and can only be anticipated to a limited extent by testing companies or guidebook publishers. Similarly, in mathematics and science education, it usually is inappropriate to pretend that students can be put in order from "worst" to "best" without recognizing the facts such as: (a) all students develop complex profiles of strengths and weaknesses[1]; (b) different jobs often demand quite different profiles of understandings and abilities; (c) a person who has excellent credentials for one job may be ill suited for another; (d) progress can be made along a variety of paths; and (e) the significance of achievements often depend on the conditions under which they occur.

The first table in this section suggests that the kind of operational definitions that will be emphasized in this chapter should be less concerned with making value judgements in the context of artificial ordeals. Rather, they should focus on finding ways to produce auditable trails of documentation that yield assessments of productivity which are based on representative samples of closely inspected student work. Table 6.2 also emphasizes the fact that productive operational definitions should lead to useful information for decision makers.

---

[1] An algebraist is not necessarily a good topologist. A topologist is not necessarily a good tax consultant. A tax consultant is not necessarily a good programmer. A person who is good at following rules is not necessarily a person who is good at creating new sets of rules (when new programming languages need to be created). A person who is good at working along to produce answers to other peoples' well framed questions is not necessarily a person who is good a designing solutions to problems in which teams of specialists must work together, and there is no single "correct" ways to balance conflicting factors such as cost, quality, speed, and simplicity.

TABLE 6.2

Distinctions Between Traditional Testing and Representative Samples of Assessed Work

| *Traditional Tests* | *Representative Work Samples* |
|---|---|
| Generally focus on:<br>(a) A small number of high-stakes but low-interest situations.<br>(b) Short-answer questions that emphasize a large number of low-level facts & skills. | Generally focus on:<br>(a) A larger number of low-stakes but high-interest<br>(b) Project-sized activities emphasizing smaller numbers of deeper & higher order understandings and abilities. |
| *Summative Information*<br>Focuses on decision-making issues of administrators & policy makers. | *Formative Information*<br>Focuses on decision-making issues of students and teachers. |
| *Pretest–Post Designs*<br>Are intended to prove that "it" works; but, they often promote conditions which minimize the chances of success by measuring progress in terms of deficiencies with respect to simplistic "check list" of benchmarks (facts, skills, behaviors). | *Dynamic Feedback*<br>Focuses on monitoring and documenting progress while also encouraging development in directions that are increasingly "better". . . without using simplistic (mechanistic) definitions of "best" to determine next steps. |

The following kinds of questions shouldn't be overlooked when judgements are made about the relative strengths and weaknesses of alternative forms of assessment (and operational definitions of goals). Who are the most important decision makers whose needs should be addressed? What are the priority-type decisions they must make? What kind of information is needed to make wise decisions? Is rapid feedback more important than accuracy? Is it sufficient to reduce all information to a single-number score? Or, is it important to see details about profiles of strengths and weaknesses. Is it sufficient to receive information about groups? Or, is it important to get details about individuals? Is it sufficient to get information only infrequently, and to generate reports only after long delays? Or, is it important to get rapid feedback so strategic improvements can be made?

Instruments that address some of the preceding needs do not address others. So, a "one size fits all" strategy is not likely to be appropriate for all decision makers. Also, it is not necessary in an age when most modern decision makers have computers sitting on their desks. Even decision makers who don't have access to computers are familiar with modern newspapers in which nearly every section from business to sports includes a complex display of computer-generated information that is graphic, multidimensional, and (perhaps) interactive. Consequently, in such a society, technology makes it possible to gather information from multiple sources, to display the results for multiple purposes, and to produce reports that are simple without being simplistic. In particular, there is no need to simplify information only by relying on one-number characterizations of students, teachers, programs, or schools.

Another distinction that is relevant to our discussions is highlighted in the cartoonist Gary Larson's book "Night of the Crash Test Dummies" (1988). That is, in many everyday situations when we test something, we don't care whether the thing being tested is being injured or destroyed in the process. But, in education, in general, we *do* care. This is why, educators usually should obey a basic principle that was emphasized long ago in the preamble to the Hippocratic Oath in medicine. That is: *First, don't do harm!*

The preceding point is especially important in mathematics and science education because the need for severe artificial constraints on time and resources usually causes traditional standardized tests to restrict their attention to extremely brief and superficial work samples, as well as to equally simplistic scoring procedures. Consequently: (a) these tests tend to focus on shallow, narrow, and biased subsets of the understandings and abilities that are needed for success in more representative samples of work (Resnick & Resnick, 1993); (b) the understandings and abilities that they assess tend to be poorly aligned with the deeper and higher order knowledge emphasized in high quality curriculum standards (Romberg, et al., 1991); and (c) they seldom give adequate attention to the most important achievements of outstanding students, teachers, and programs. In other words, standardized tests tend to emphasize what is *not* important, while neglecting what is important. Therefore, when scores on such tests are used to predict performance in more realistic problem solving situations, they generally account for far less than 20% of the variance (Willingham, et al., 1983; Gardner, 1993a; Goleman, 1997). Yet, such tests continue to be used as de facto "operational definitions" to specify goals of instruction. Why?

Even testing companies don't use standardized tests to make high stakes decisions about employees' promotions and salary increases, or about the productivity of various components of their businesses. This is because many other cost-effective ways usually exist to produce auditable trails of documentation that provide information about achievement and productivity. Yet, in education, in spite of obvious alternatives to testing that are familiar in other areas of life, it is common to assume that expensive testing programs offer the only possible way to hold students, teachers, and schools accountable for high achievement.

At the same time that standardized tests are touted as the best way to emphasize accountability, and to encourage students, teachers, and schools to take responsibility for their own achievements, they actually provide ideal ways to *avoid* responsibility. This is because they rarely provide feedback that is in time or on target to do anything about the results. For example, unlike professional people in real life job situations, students are encouraged to imagine that high scores on a college entrance examination will relieve them of the responsibility to produce other forms of credible evidence showing productivity in challenging situations over long periods of time. Or, schools imagine that, by claiming to teach to tests, they can avoid responsibilities to provide better courses, better counseling, and better support to actually help students to get admitted (and to be successful) in excellent colleges and professions.[2]

---

[2] It is misleading to speak of treating students fairly on a test (or in a textbook or teaching program) if the test as a whole reflects obsolete, superficial, and instructionally-unproductive biases about the nature of mathematics, problem solving, teaching, and learning. Yet, if a narrow conception of mathematical ability is correlated with a more representative interpretation, then testing specialists often treat modest (.38) correlations as if they are sufficient (i.e., sufficient for selecting small-but-adequate numbers of students for access to scarce-but-adequate resources). Yet, today, even beyond concerns about fairness and equity, national assessment priorities have changed. At a national level, our foremost problem is not to *screen* talent; it is to *identify* and *nurture* capable students. The pool of students receiving adequate preparation in mathematics is no longer adequate; and, far too many capable students are being "shut out" or "turned off" by textbooks, teaching, and tests which give excessive attention to views of mathematics and problem solving that are fundamentally inconsistent with national curriculum and assessment standards for school mathematics and science.

The most serious shortcomings of standardized tests tend to occur when they go beyond being used as *indicators* of achievement to being used as if they *defined* what we mean by achievement. One argument in favor of teaching-to-tests is based on the fact that international assessments of students' mathematical abilities continually show that students in the United States rank far below their counterparts in other countries with prosperous economies. Yet, there is no country that has a stronger tradition of teaching to tests than the United States; and, the United States spends orders of magnitude more money on standardized testing than any other country. This is why there is no other country that has anything like the testing industry that exists in the United States. Consequently, while the United States is preoccupied with trying to catch up in low level "basics" emphasized on tests, many competing countries are shifting toward "thinking about complexity" and toward preparation for success in a technology-based society. Implications of these trends are clear. If the USA catches up, it will once again be left behind!

It is well known that pressures to teach to tests often exert strong forces on schools to "dumb down" their curriculum offerings in order to emphasize the simplistic conceptions of expertise that standardized tests embody. For example, consider the Educational Testing Service's to Scholastic Aptitude Test (SAT) that is used to select students for admission to many colleges.[3] It has been a long-standing ETS policy that the SAT should include no

---

[3]Both of the authors of this chapter have served as deans as well as faculty members working with admissions officers at a variety of highly selective universities in the United States and Australia. Consequently, we've learned that, contrary to popular beliefs of those not directly involved in the decision making process, scores on standardized tests tend to be given far less attention in admissions decisions than factors such as: rank in class, teachers recommendations, recommendations from credible alumni, consistent records of good grades in challenging courses in schools whose programs are known to be strong, and other evidence of productivity and diversity. For example, at Purdue University, the admissions office clearly states that scores on standardized tests are considered only after at least six of the preceding kinds of factors have been assessed—and, even then, test scores are used mainly to grant admissions to students who are able to provide little other evidence of their capabilities. Unfortunately, many other colleges who follow similar policies are less forthright in acknowledging that their decisions are based on professional judgements rather than on rigid decision formulas. That is, to avoid the responsibility

(continued)

content that goes beyond ninth grade mathematics. Furthermore, SAT only claims to be measuring general aptitude rather than achievement. That is, it claims to be measuring "intelligence" that is presumed to be uninfluenced by particular instruction. Therefore, teaching to this test implies encouraging students to focus on mathematics that never goes beyond the ninth grade level; and, it also implies neglecting deeper and higher order thinking that is tied to the use of powerful and highly content-specific tools such as those needed for success in a technology-based society.

The preceding facts are among the most obvious reasons why there is only an unimpressive .38 correlation between SAT scores and success in introductory college courses—even when the only college courses that are considered are those in which students' grades are based on nothing more than performance on easy-to-score tests (Willingham et. Al., 1990). In other words, standardized tests don't even do an especially impressive job of predicting performance on freshman tests; and, correlations fall toward zero if success in post-graduation work experiences is considered—or if judgements about success are based on work samples that require more complex forms of thinking than those emphasized on standardized tests (Goleman, 1997; Shavelson & Baxter, 1992).

Later in this chapter, other explanations will be given about why teaching to tests, over a long period of time, tends to be a losing strategy. For now, a brief explanation is that: (a) students soon forget disorganized lists of facts and skills; (b) when facts and skills are "mastered" one-at-a-time and in isolation, students may never learn when to chose which one to use in particular situations; and (c) when instruction emphasizes only facts and skills, or content-independent problems solving processes, other exceedingly important goals of instruction are ignored.

---

of informing students that they didn't get admitted because they had poor records of performance in weak programs, it is easier and less risky to simply lead people to believe that they were rejected because of low test scores.

## RELATIONSHIPS BETWEEN MEASUREMENT INSTRUMENTS AND OPERATIONAL DEFINITIONS OF UNDERLYING CONSTRUCTS

Why is it so common, in education, to assume that *indicators* of achievement must be used to *define* goals of instruction? Thermometers measure temperature; but, this doesn't lead us to believe that simply causing the mercury to rise will do anything significant to change the weather.

Relationships between measurement, meaning, and dictionary-style definitions can be complex. Yet, these complexities need not lead to confusion about differences between measurement instruments and the constructs they are intended to measure. For example, consider the case of clocks and wrist watches. They are used to measure time; yet, even today, many basic issues remain unresolved about the measurement and meaning of time. In fact, during the past decade, many of the world's leading scientists have written books that attempt to explain various ways of thinking about the nature of time (Aveni, 1989; Coveney & Highfield, 1990; Davies, 1995; S. J. Gould, 1988; Hawking, 1988; Sachs, 1987; Shlain, 1991; Thorne, 1994). Yet, even though many deep issues remain unresolved about the nature of time, most ordinary people believe that they understand the concept of time sufficiently well simply because, for most practical purposes, they know how to measure time and they know how to deal with its effects.

Why do confusions persist in education about relationships between our ability to *measure* a construct and our ability to *define* it? Some reasons are because, in education: (a) it is common to treat means-to-ends as if they were ends-in-themselves; (b) it is common to treat effects as if they were causes; and (c) it is common to treat the consequences of new knowledge as if these consequences were the most essential parts of the underlying knowledge itself. In particular, high scores on tests often are treated as if they went beyond being *indicators* of achievement toward actually *being* the goals of instruction. If similar lines of reasoning were used in fields such as medicine, we would treat symptoms as if they were the underlying diseases.

The preceding fact is especially important in mathematics and science education because many of the most important constructs that students should learn are not reducible to facts and skills (*behavioral objectives* of instruction) nor to global processes (*process objectives* of instruction). Instead, they represent a third type of instructional goal that modern cognitive

psychologists might refer to as *cognitive objectives* of instruction. They are models, metaphors, or other complex conceptual systems that are used to describe, to explain, or to construct complex systems in real or imagined worlds. For example, in science education, children develop their own ways of thinking about electricity, heat, light, magnets, and many other phenomena that they encounter in everyday situations; and, changing students' ways of thinking usually must involve more than simply forcing them to learn a few new facts or skills. Unless previously developed ways of thinking are challenged and tested, they tend to persist. For instance, if a child explains electricity as if it were like water running through tubes, then new facts and rules that are memorized are incorporated into these existing ways of thinking—rather than leading to new models, metaphors, and conceptual systems.

To illustrate the way that holistic conceptual systems resist change, there is perhaps no better example to consider than the way many educators think about assessment. In past stages of history, educators have tended to think about the mind (and about the nature of mathematical knowledge) as if it were similar to the most sophisticated technology of the preceding age. For example, as civilizations evolved from the industrial revolution through the electronics revolution to the current age of biotechnologies, educators have shifted from machine metaphors (based on hardware) to computer metaphors (based on software) to organic metaphors (based on wetware of the type that characterizes the processes of the human brain). Yet, in the area of assessment, simple input–output models continue to dominate that are based on machine metaphors. Consequently, to improve the assessment practices of schools and other educational agencies, it is not enough to introduce better test items. An entire paradigm shift will be needed which fundamentally rethinks issues ranging from data collection, data interpretation, data analysis, and the nature of reports (Lesh & Lamon, 1992).

Similar observations have been made about other components of curriculum reform, such as those that focus on teacher development, curriculum development, or program development. For example, to explain why piecemeal approaches to curriculum innovation are seldom effective, a report from the National Council of Teachers of Mathematics makes the following observation.

> In spite of the best intentions of developers and implementors, it was unreasonable to expect that new products or programs would be used as intended in most schools and classrooms. The reason for this is that public schools as they now operate are integrated social systems. Tinkering with parts, such as changing textbooks or the number of required courses, fails to change other components of the system. The traditions of the system force new products to be used in old ways. Current educational

> practice is based on a coherent set of ideas about goals, knowledge, work, and technology that came from a set of "scientific management" principles growing out of the industrial revolution of the past century. These ideas about schooling need to be challenged and replaced with an equally coherent set of practices in light of the economic and social revolution in which we are now engaged. Current school mathematics operates within a coherent system; reform will happen only if an equally coherent system replaces it. (National Council of Teachers of Mathematics, 1989)

According to ways of thinking borrowed from the industrial revolution, teachers have been led to believe that the construction of mathematical knowledge in a child's mind is similar to the process of assembling a machine, or programming a computer. That is, complex systems are thought of as being nothing more than the sums or their parts; each part is expected to be taught and tested one-at-a-time, in isolation, and out of context; and, the parts are assumed to be defined operationally using naive checklists of condition-action rules.

In contrast to the preceding perspective, scientists today are investigating complexity theory where the processes governing the development of complex, dynamic, self-organizing, and continually adapting systems are quite different than those that apply to simple machines. Parts interact. Logic is fuzzy. Whole systems are more than the sums of their parts; and, when the relevant systems are acted on, they act back.

Piaget was one of the earliest and most prominent researchers to reveal the nature of some of the most important conceptual systems that students must develop to make judgements about quantities ranging from simple counts, to lengths, to areas, to speeds, to forces, to probabilities (Piaget & Beth, 1966; Piaget & Inhelder, 1958). Similarly, during the past 30 years, some of the greatest achievements of mathematics and science education research have been to generate detailed information about what it means to develop deeper or higher order understandings of the conceptual systems that underlie students' developing knowledge about constructs ranging from fractions (Behr, Harel, Lesh, & Post, 1991) to forces (Hestenes, M. Wells, & Swackhamer, 1992) to functions (Kaput & Nemirovsky, 1995), to frequencies (Gal & Garfield, 1997).

The preceding increases in knowledge have been primary factors leading to the successful development of curricula standards produced recently by organizations such as the National Council of Teachers of Mathematics (1989), the American Association for the Advancement of Science (1993), and the National Research Council-National Academy of Sciences (National Research Council, 1996). Yet, in spite of enormous progress clarifying the nature of students' developing mathematical and scientific knowledge, many leaders in research have

BEST COPY AVAILABLE

been reluctant to specify ways to measure what they mean by "understanding" for relevant conceptual systems. Why? One reason is because of fears that educators will confuse measurement instruments with the constructs they are intended to assess; but, another reason is because, when mathematics and science educators describe what kinds of understandings and abilities should be stressed in instruction, these curriculum standards for *instruction* do not necessarily provide straightforward performance standards for *students* (Romberg, et al., 1991). For example, saying that mathematics instruction should stress connections among ideas does not make clear how "connectedness" can be assessed in a student's thinking. Or, saying that flexibility of thought should be encouraged in instruction does not make it clear how flexibility could be assessed in ways that are useful, reliable, and unbiased. Similarly, specifying what we mean by terms like *understanding* tends to become especially problematic when modifiers are added such as: *concrete* understanding, *abstract* understanding, *symbolic* understanding, *intuitive* understanding, *situated* understanding, *higher order* understanding, *instrumental* understanding, *relational* understanding, *deeper* understanding, or *shared* understanding. In fact, even in the case of familiar terms such as *teaching*, *learning*, and *problem solving*, definitional difficulties often arise when modifiers like *good* are added. What do we mean when we speak of good teaching, good problem solving, good assessment tasks, or good instructional activities?

So far in this chapter, we've seen that the relationships often are not straightforward between measurement instruments and operational definitions of underlying constructs, or between curriculum standards for *instruction* and performance standards for *students*. The next section will describe some alternative types of operational definitions that are familiar in fields outside education, and that may provide useful ways to deal with the preceding kinds of complexities.

## ALTERNATIVE TYPES OF OPERATIONAL DEFINITIONS FOR SPECIFYING GOALS OF INSTRUCTION

In sciences ranging from physics to astronomy to biology to geology to anthropology, many levels and types of operational definitions are commonly employed to deal with constructs such as neutrinos, strange quarks, and other subatomic particles that have characteristics with whimsical names like *charm* or *flavor*. Furthermore, many of these constructs have characteristics similar to those that are of greatest interest in mathematics and science

education. For example, the existence of a given entity may be inseparable from the complex holistic systems in which it is embedded. Or, the entity may not be observable directly, but rather may be knowable only by its effects.

Consider the case of the neutrino where huge vats of heavy water are surrounded by photo-multipliers in order to create situations in which the effects of neutrinos are likely to be observable and measurable. Yet, even under these conditions, neutrinos cannot be observed directly. They can be known only through their effects; and, between the beholder and the beheld, elaborate systems of theory and assumptions are needed to distinguish signal from noise, and to shape interpretations of the phenomena under investigation. Also, small changes in initial conditions often lead to large effects that are essentially unpredictable; and, observations that are made induce significant changes in the systems being observed. Consequently, both researchers and their instruments are integral parts of the systems that scientists are hoping to understand and explain.

Educators are not alone in their need to deal with constructs and phenomena that have the preceding characteristics. Furthermore, in sciences where fanciful terms such as *charm* and *flavor* are used to refer to characteristics of inanimate objects, it is clear that educators calling themselves constructivists are not alone in their recognition that reality is a construct. Throughout the physical sciences, life sciences, and social sciences, many of the most important "entities" that we need to understand and explain are products of human construction.

When devices such as cloud chambers and cyclotrons are needed to record and measure illusive constructs, it is clear that being able to measure a construct does not guarantee that a corresponding dictionary-style definition will be apparent. Conversely, even when scientists can give a dictionary-style definition of a construct (such as a black hole in astronomy), this doesn't guarantee that procedures will be clear for observing or measuring the construct.

In general, in the physical sciences, to develop useful operational definitions of complex phenomena (photons, neutrinos, etc.), a minimum requirement is that explicit procedures must be described specifying: (a) how to create a situation in which the relevant construct will occur; (b) how to detect its effects and sort out relevant from irrelevant information; and (c) how to make comparisons among various observed occurrences. Therefore, such operational definitions involve three parts that are similar, in some respects, to the following three parts of traditional types of behavioral objectives in mathematics and science education.

Behavioral Objectives Involve Three Parts

> Given {specified conditions} the student will exhibit {specified behaviors} with identifiable quality {perhaps specified as percents correct on relevant samples of tasks, or perhaps specified in terms of a correspondence with certain criteria for excellence}.

In mathematics and science education, the main problem with behavioral objectives is that not all forms of learning consist of rules (facts, skills, procedures); and, if attempts are made to reduce more complex conceptual systems to checklists of rules, the following sorts of difficulties arise. Even though students (teachers, programs, or schools) who develop a given conceptual system (C) may be expected to exhibit certain behaviors (B1, B2, B3, . . . Bn) on tasks (T1, T2, T3, . . . Tn), it usually is possible for these problem solvers to exhibit behaviors B1, B2, B3, . . . Bn *without* developing the underlying conceptual system C. This is because the relationship between C and its various manifestations B1, B2, . . . Bn tends to be associative rather than causal, and symptomatic rather than definitive.

In general, the development of complex conceptual systems is similar to the development of complex and dynamic systems that occur in other types of situations (such as sports, arts, or business). For example, coordinated and smoothly functioning systems usually involve more than the simple sums of their parts. In fact, characteristics of the parts tend to be inherited from characteristics of the system as a whole at least as much as the converse is the case. For instance, it may be true that a great artist (or a great athlete) should be able to perform well on certain basic drills and exercises (calisthenics); nonetheless, a program of instruction (or assessment) that focuses on nothing more than these checklists of basic skills is unlikely to provide adequate approaches to excellence. For example, if we taught (and tested) cooks in this way, we'd never allow them to try cooking a meal until they memorized the names and skills associated with every tool at stores like Crate & Barrel and Williams Sonoma. Or, if we taught (and tested) carpenters in this way, we'd never allow them to try building a house until they memorized the names and skills associated with every tool at stores like Ace Hardware and Sears. Yet, whereas such an approach clearly is foolish in fields like cooking, carpentry, art, or sports, it is employed routinely in mathematics and science education. If fact, when students' performance deteriorates after schools focus instruction on a checklists of low level skills, a common reaction is for schools to focus on an even more detailed checklist. In fields such as cooking, carpentry, or sports, such an approach would be

recognized as being the disease for which it purports to be the cure! But, in education, such approaches persist.

The next section of this chapter will describe some additional relevant shortcomings associated with the use of behavioral objectives in mathematics and science education. But, in this section, the goal is to describe alternatives to behavioral objectives as means of defining observable and assessable goals of instruction. In past publications, we have referred to some of the most productive of these new types of operational definitions as "cognitive objectives" of instruction (Lesh & Lamon, 1994). This name seems appropriate because they focus on students' *interpretations* of situations rather than focusing on their *actions* in these situations.

Whereas behavioral objectives collapse three different kinds of statements into a single condition-action rule, cognitive objectives function similarly to the ways cyclotrons, cloud chambers, and vats of heavy water are used in physics. That is, they are defined operationally by specifying: (a) *situations* that optimize the chances that the targeted construct will occur in an observable form; (b) *observation tools* that enable observers to sort out signal from noise in the results that occur; and (c) *quality assessment criteria* that allow meaningful comparisons to be made among alternative possibilities. For example:

> *Situations*: At the end of this chapter, Appendix A gives an example of a thought-revealing activity in which three-person teams of middle school students confront the need to develop weighted sums, weighted averages, or some other construct to aggregate information of qualitatively different types (such as times, distances, rankings, or qualitative information that must be quantified). Other examples are given in Chapters 21 and 23 of this volume. For instance, the tasks called "The Sears Catalogue Problem" (Chapter 21, this volume) and "The Summer Jobs Problem" (Chapter 23, this volume) both require students to deal with situations in which they must quantify relationships among quantities, then aggregate these relationships into a single measure.

For the purposes of this chapter, the main point that we want to emphasize about the preceding types of activities is that they have characteristics similar to those that apply to cloud chambers or other devices that are used to observe or measure subatomic particles. For example, the relevant construct is *elicited by* the task description or the device, but it does not *reside in* the task description or the device. In particular, for the kind of thought-revealing activities that we emphasize, the mathematical constructs do not reside in the problem



BEST COPY AVAILABLE

statements, and hence they are not task variables; instead, they reside in the responses that students generate, and hence might be called response variables.

*Observation Tools*: In mathematics education, the kind of thought-revealing activities that we describe in Chapters 21 and 23 of this volume are examples of tasks that require students to reveal their ways of thinking. This is because the products that students develop include more than calculated results consisting of a single number answer; they also involve descriptions, constructions, explanations, or justifications that may be embedded in spreadsheets, graphs, equations, and other representational systems. Consequently, as students goal mathematize (quantify, coordinatize) the problem solving situations, the explicit mathematical interpretations that are generated reveal significant information about what kind of mathematical relationships, patterns, and regularities the student is taking into account. Furthermore, because the descriptions, constructions, explanations, and justifications that students produce are not merely parts of the processes of producing answers to questions, but are essential parts of the products themselves, it sometimes is possible for these products to reveal significant information about the processes that produced them. Also, because these products often need to be shareable, modifiable, transportable, and reusable, students often must go beyond thinking with a given construct to also think about it; and, they also must think about them in a way that generalizes beyond the specific situation that is given. Thus, it often is straightforward to deal with deeper and higher order understandings and abilities. Nonetheless, even if the products that students produce explicitly reveal significant aspects about the ways of thinking that they are using to make sense of the situation, various levels and types of products nearly always are possible, and it is not necessarily a straightforward task to analyze and interpret their results. For example:

- It nearly always is possible for students' responses to make use of a variety of representational systems (involving written symbols, spoken language, graphs or graphics, diagrams, and concrete models).
- It nearly always is possible for several, equally acceptable responses to be based on different decisions about trade-offs involving factors like accuracy, precision, simplicity, or timeliness.

To deal with the preceding kinds of complexities, teachers participating in the authors' past research projects have found it useful to develop "Ways of Thinking Sheets" that sometimes resemble large posters displaying snippets from the products that students produce for a given problem (Katims, Lesh, Hole, & Hoover, 1994a). For example, for the "Softball Problem" that is given in Appendix B, some relevant snippets might show several ways in which average-ability, middle school students often develop creative ways to: (a) quantify qualitative information (such as the weights or values that they assign to various categories of information given in the problem); (b) combine information involving diverse, qualitatively different, types of information (entailing ranks, distances, times, or value points); (c) make their assumptions explicit, and allow modifications to be made to deal with alternative assumptions; and (d) use a range of graphs, equations, or other representations. Other "Ways of Thinking Sheets" are described in Chapter 21 of this book.

*Quality Assessment Criteria*: For the thought-revealing activities that we describe in Chapter 21 of this book, the constructions, descriptions or explanations that students produce provide conceptual tools that are requested by a specified client and for a specified purpose. Furthermore, the client and purpose are specified in such a way that this information provides straightforward criteria for making judgments about the strengths and weaknesses of alternative approaches. Nonetheless, to assess final results that are produced by different teams of students, teachers (and students) generally find it useful to use guidelines for assessing the usefulness of results. An example set of guidelines that can be used for the "Softball Problem" is given in Appendix B at the end of this chapter.

Note: It has become common, in projects emphasizing performance assessment, to use the term *scoring rubric* to refer to guidelines for assessing the quality of students' work. But, we prefer to avoid this term because many scoring rubrics that have been developed operate on principles that are aligned more closely with checklists of behavioral objectives than with the kind of cognitive objectives that we wish to emphasize. Instead, we emphasize procedures similar to those used in Olympic competitions (basketball, gymnastics, and equestrian events), performing arts or crafts (dancing, playing musical instruments, and building houses), or businesses (assessing productivity for individuals, groups, and sites). In such contexts, the goal of assessment often involves complex systems; and, to assess these systems, the procedures that are used often involve three parts similar to those being described

here. That is they involve creating: (a) situations in which these systems will be elicited and observable; (b) tools to help recognize important characteristics of the relevant systems that occur; and (c) tools to assess the relative strengths and weaknesses of the learner's systems that are elicited. Frequently, in these situations, the act of assessing complex performances is itself a complex performance. But, this need not be the case. For example, in complex businesses, or in sports or arts where complex performances are emphasized, final tests of success often are measured by scoring points, by making money, or by winning against an opponent. Yet, in the preceding situations, performers seldom do well if they are too preoccupied with only scoring points, making money, or winning. Because it is clear that the goal is to develop complex systems, there is no confusion between the goal and the criteria for assessing the goal. So, participants are not likely to imagine that complex performances can be reduced to a checklist of simple behaviors.

Taken together, the preceding kind of situations-plus-observation-tools-plus-quality-assessment-criteria often provide useful operational definitions of complex systems. This is because they include explicit procedures for describing: (a) how to create situations in which the relevant constructs will be likely to occur; (b) how to recognize and describe the relevant constructs when they occur; and (c) how to make comparisons among the alternatives that occur. In particular, in the case of conceptual systems that students develop during the solution of individual problem solving sessions: (i) model-eliciting activities put students in situations where they confront the need to produce a given type of construct, and where the products that they generate require them to reveal explicitly important characteristics of their underlying ways of thinking; (ii) ways of thinking sheets focus on ways of recognizing and describing the nature of the constructs that students produce; and (iii) guidelines for assessing the quality of students' work provide criteria that can be used to compare the usefulness of alternative ways of thinking.

## COMPLEX SYSTEMS HAVE DYNAMIC MULTIDIMENSIONAL PROFILES OF COMPETENCE

To identify other issues that are important to consider when operational definitions are formulated, it is useful to recognize that, in mathematics and science education, three-part operational definitions are especially useful because many of the "entities" to be observed and

assessed involve complex, interacting, self-organizing, and continually adapting systems that cannot be reduced characterizations based on a checklist of condition–action rules. These complex systems include cognitive structures (which underlie a given student's understanding of a particular idea), conceptual systems (which underlie a student's understandings of relationships among ideas in a given topic area), biological systems (involving children's development), and social systems (within classrooms and schools).

Regardless of whether the individuals we want to assess are students, groups, teachers, classes, programs, or schools, the following two characteristics usually apply to the individual we want to investigate.

1. In general, there is no single "best" type of individual; every individual has a complex profile of strengths and weaknesses; and, individuals who are effective in some ways and under some conditions are not necessarily effective in others.

2. Individuals at every level of expertise must continue to adapt and develop; there exists no fixed and final state of excellence; in fact, conceptions of excellence tend to be among the most important characteristics that evolve during the development of individuals.

These two characteristics have important implications for the types of operational definitions that need to be emphasized in mathematics and science education. For example:

> *For Programs*: When evaluating large and complex program innovations, it is misleading to label them "successes" or "failures" (as though everything that successful programs did was effective and everything that unsuccessful programs did was not effective). All programs have profiles of strengths and weaknesses. Most of them work some of the time, for some students, in some ways, and in some circumstances, but none of them work all of the time, for all students, for all purposes, and in all circumstances. Furthermore, success depends generally on how much and how well the program is implemented. If only half of a program is implemented, or if it is implemented in only a half-hearted way, then complete success is not expected. Also, as soon as a program begins to achieve success, fundamental changes tend to be induced into the situation where they are expected to function, and new rounds of adaptation tend to be needed. Therefore, when programs are observed and assessed, the kinds of descriptions that are needed should be multidimensional, conditional, and continually evolving.

*For Teachers*: As teachers develop, they frequently notice new things about their students, about their instructional materials, and about the ideas and abilities that they are trying to help their students learn. Consequently, these new observations often create new needs and opportunities that, in turn, demand additional development. Current capabilities be characterized by a single point on a "good–bad" scale; profiles of strengths and weaknesses always can be identified; and, the conditions in which these profiles have been achieved must be taken into account. For instance, no teacher is equally effective for all grade levels (kindergarten through calculus), for all types of students (handicapped through gifted), and for all types of settings (inner-city to rural). No teacher can be expected to be "good" constantly in "bad" situations; not everything that experts do is effective, nor is everything that novices do ineffective. Characteristics that lead to success in one situation (or for one person) often turn out to be counterproductive in other situations (or for another person). Finally, even though gains in students' achievement should be one factor considered when documenting the accomplishments of teachers (or programs), it is foolish to assume that the best teachers always produce the largest student learning gains in students. What if a great teacher chooses to deal with only difficult students or difficult circumstances? Clearly, expertise in teaching is plural, multidimensional, nonuniform, conditional, and continually evolving.

*For Students*: Chapter 23 of this book (Lesh & Lehrer) gives examples to show how students' responses to model-eliciting activities involve the development of complex descriptions or explanations. Furthermore, it also shows how these interpretations tend to develop through a series of modeling cycles in which the results of each cycle enable students to take into account new and different kinds of information, which, in turn, creates the need for additional refinements or embellishments in the underlying interpretation. In such problem solving situations, the development of students' ways of thinking tends to involve sorting out unstable conceptual systems at least as much as it involves assembling stable systems; and, as more sophisticated interpretations evolve, students often go beyond thinking *with* a given construct to thinking *about* the construct (e.g., by identifying assumptions that have been made and by identifying alternatives that might be appropriate under a variety of conditions).

> Also, development tends to occur along a variety of relatively independent dimensions such as concrete–abstract, intuitive–formal, or situated–decontextualized. So, again, improvements tend to be multidimensional and unbounded, and, in general, the evolution of these systems obeys rules similar to those that apply to other complex, self-regulating, and continually adapting systems.

Three final characteristics should be mentioned that pertain to operational definitions involving the development of students, teachers, and programs. First, the development of these problem solvers tends to be highly interdependent. Even though, for some purposes, it may be possible to deal with the development of students, teachers, and programs as if they were isolated and independently functioning entities, in reality, they tend to be interdependent and indissociable. Second, when something (or someone) acts on any one of these complex systems, they tend to act back. Therefore, development is not solely the result of the passive acceptance of information. Third, researchers (as well as the instruments that they use) usually are integral parts of the systems that they are hoping to understand and explain. Therefore, observations often induce significant changes in the systems being observed, and small influences often lead to large effects that are essentially unpredictable. In other words, investigations involving the interdependent development of students, teachers, and programs involve many of the same characteristics that occur when complex and continually adapting systems are investigated in fields like physics, geology, astronomy, and biology. These characteristics include the following:

> *Selectivity*: Observation involves an act of selection to distinguish signal from noise, to distinguish significant patterns from meaningless perturbations, and to select a productive level of detail to guide analysis. Therefore, different observers may emphasize or omit data (actions or contextual details) that are viewed as significant by participants, by other observers, or from other perspectives.

> *Indeterminacy*: Observations often have a powerful influence on the phenomena being observed, and the only way to observe certain "objects" may be to induce changes in them and to observe their interactions with other phenomena.

*Subjectivity*: Observers cannot claim to represent the meaning that the activity has for the observed participant(s), only the meaning that the researcher-observer has developed.

*Reductionism*: The systems being investigated are not completely decomposable; that is, the systems as a whole are more than the simple sums of their parts, and characteristics of the parts are influenced by characteristics of the system as a whole.

*Decontextualization*: When a given system is isolated from other relevant systems and from the contexts that led to its development, some of its characteristics may be changed in significant ways.

*Reification*: Observations often accord a reality to the "observables" that conceals the inferential chain by which the researchers build representations of the relevant constructs.

*Dissociation*: Research observations occur within a larger social event (the research enterprise as a whole), which has its own participants (observer and observed), context (immediate social setting and surrounding research setting), purpose (observed task and overarching research aims), actions (including those of the researcher), and consequences.

## MULTITIERED TEACHING EXPERIMENTS PROVIDE MULTIDIMENSIONAL INTERACTING TRACES OF DEVELOPMENT

In Chapter 9 of this book, Lesh & Kelly describe techniques that were designed especially to investigate the multidimensional development of complex, interacting, self-regulating, and continually adapting systems (such as those associated with the models and conceptual systems that are used by students, groups, teachers, classes, schools, or school systems to structure their experiences). Consequently, multitiered teaching experiments were designed to investigate situations in which the challenges described in the previous section occur with regularity. Also, because multitiered teaching experiments focus on cognitive objectives of instruction, they provide alternatives to pretest–posttest designs for assessment, where the

goals of instruction tend to be restricted to checklists of behavioral objectives. For this reason, it is possible for multitiered teaching experiments to go beyond being research designs for *observing* development; they also can be used as assessment designs for *documenting* development, and, at the same time, they can be used as instructional designs for *stimulating* and *facilitating* development. Within such a framework, optimization and documentation are not incompatible processes. In fact, the reflexivity of assessment and instruction is an analogue of the self-adaptive behaviors displayed by many living things that are capable of optimizing their performance in response to changing circumstances.

A typical multitiered teaching experiment involves three parallel and superimposed studies into the interacting development of students, teachers, and programs. The student-level study might be a 15-week investigation about the nature of students' developing ways of thinking about specially selected sequences of thought-revealing activities (Chapter 21, this volume). The teacher-level study might use the preceding tasks for students as the basis for teacher-level activities in which it is possible to investigate the development of teachers' integrated, mathematical, psychological, and pedagogical ways of thinking about students' evolving learning and problem solving abilities. Finally, the program-level study might use the tasks for students and teachers as contexts for investigating the development of relevant ways of thinking for administrators, parents, and policy makers concerning both the nature of their goals of instruction, and concerning the nature of high-quality teaching and learning environments.

Whereas thought-revealing activities for students are designed especially to reveal important characteristics of students' ways of thinking, multitiered teaching experiments also need to employ teacher-level activities in which teachers continually reveal, test, and refine or revise their ways of thinking. Examples of such teacher-level activities include generating: (a) observation sheets that colleagues could use to make significant observations about students as they are working in groups; (b) "ways of thinking sheets" that colleagues could use to give feedback to students about the strengths or weaknesses of their work; and (c) quality assessment procedures that colleagues could use to lead discussions with students aimed at assessing the quality of the products that students produce.

For the purposes of this chapter, three of the most important characteristics of multitiered teaching experiments are that:

1. They use formative feedback and consensus-building to provide conditions that make it possible for students, teachers, and/or programs to develop in directions that are continually "better" without basing the next steps on preconceived notions of "best."

2. They emphasize the use of self-documenting activities that encourage students, teachers, and/or programs to learn while simultaneously producing trails of documentation that reveal important characteristics about the nature of what is being learned.

3. The preceding trends allow inferences to be made about future developments that are likely to occur.

In other words, multitiered teaching experiments generate information that includes more than unorganized collections of isolated data points. The results also include traces of interactions, and these trajectories often lead to multidimensional descriptions of developing expertise.

One of the most important assumptions underlying the teacher level of multitiered teaching experiments is that, to improve teachers' teaching practices, it is not enough to ensure familiarity with a checklist of behavioral objectives; teachers also need to develop productive ways thinking about their teaching experiences. In particular, teachers need to develop productive ways of thinking about their students' learning and problem solving experiences.

## IN MULTITIERED TEACHING EXPERIMENTS, THE NATURE OF DEVELOPING ENTITIES EMERGES RETROSPECTIVELY

Cognitively guided instruction is a name that's been given to an approach to teacher development that focuses on helping teachers become "reflective practitioners" by becoming familiar with new insights about the nature of students' developing mathematical knowledge (T. Carpenter & Fennema, 1992). To implement cognitively guided instruction, one difficulty that arises is that, for nearly any given topic area in the K–12 mathematics or science

curricula, too little is known about the nature of students' developing knowledge, and too little is known about the types of understanding that represent expert teacher-level knowledge. For example, for topics dealing with fractions or ratios, teachers surely need to develop deeper and higher order understandings than those needed by their students. But, what is the nature of such deeper and higher order understandings? It's clear that, in addition to understanding how relevant ideas develop psychologically (in the minds of students), teachers also benefit from knowing how the ideas developed historically (why did our ancestors need this idea?), logically (what alternative ways can the truth of the idea be established?), and instructionally (what alternative approaches to the topic are offered by textbooks or other curricular materials?). But, what other understandings might be needed? For most topic areas, researchers do not know the answers to this question yet. This is why multitiered teaching experiments emphasize the use of formative feedback and consensus-building, which do not require researchers to formulate a dictionary-style definition of an expert teacher before it is possible to help teachers develop in directions that are increasingly more effective. Also, rather than *telling teachers* about their students' ways of thinking, thought-revealing activities are used so that teachers can *make firsthand observations* about their students' ways of thinking.[4]

The preceding capabilities have important implications for those who are trying to formulate productive operational definitions for their instructional goals. To see why (and how) this is true, consider an educator who says, "I don't know how to define what I mean by a great teacher, but I know some important kinds of situations to look at, and I can point to

---

[4]Teachers' observations can lead to rich portrayals (or models) of their students provided that teachers have an adequate vocabulary of constructs with which to build their evolving ways of thinking. Given this, teachers construct a bottom-up model of the student that documents and celebrates each capability (and is informed by each difficulty) rather than documenting the gap between the student's current state of knowing and an idealized goal of a state of expertise. Emphasis on this gap commits the teacher to the documentation of deficiency rather than of accomplishment. Sweller (1992) has contrasted means–ends analysis with bottom-up approaches to problem solving in favor of the latter. Sweller's argument is that schema development or conceptual connectedness is facilitated by a bottom-up approach. Teachers' cumulative construction of their pupils' capability represents an analogous form of learning.

good teaching when I see it." What such educators are saying is that they know how to create situations to elicit and assess critical aspects of teaching, but they do not know yet how to deal with the third component that is needed to generate a three-part operational definition. That is, they do not know how to formulate an adequate multidimensional description of teachers' developing knowledge and abilities. But, within the kind of dynamic feedback systems that are emphasized in our multitiered approaches to research, assessment, and instruction, the nature of developing systems emerges at the end of the experience rather than being presupposed at the beginning.

When researchers or teachers know how to elicit and assess a construct (or system), these capabilities often make it possible to create a series of construct-refinement activities that will generate progressively clearer and more useful descriptions of developments that occur. For example, in studies of developing expertise of teachers (Lesh & Kelly, 1998), it is not necessary for a given description of expertise to be locked in at the beginning of a study (and used as the basis for a pretest–posttest design). Instead, increasingly sophisticated descriptions can be refined and documented gradually over the course of the study; and, at the end of the study, the validity of the description can be based on the trajectory of development that is revealed. In fact, as teachers, programs, and schools develop, their notions of excellence in teaching are primary factors that change. Therefore, if their progress continues to be measured using conceptions of excellence that existed at the beginning of a study, this practice tends to have significant negative influences on development.

## WHAT ARE "WAYS OF THINKING" FOR PROGRAMS, SCHOOLS, OR SCHOOL SYSTEMS?

Regardless of whether researchers investigate the development of individual students, groups, teachers, programs, schools, or school systems, a fundamental assumption underlying this chapter is that it is seldom sufficient to describe increasing expertise using only a checklist of behavioral objectives. For any of the preceding kinds of problem solvers, complex systems are involved, and excellent performance on complex tasks generally involves the development of powerful models (or *ways of thinking*) that provide meaningful ways to structure these experiences. But, what does it mean to speak about the ways of thinking of a program, a school, or a school system? By ways of thinking, we mean a system's interpretation (descriptions, explanations) of experiences. How can the nature of these ways of thinking be

observed? How can their efficacy be assessed? Our answer to these questions is that procedures should be investigated that have been effective in other sciences where the systems being investigated also are complex, dynamic, and continually adapting.

Chapter 3 of this book, which is about identifying priority problems for research (Lesh & Lovitts), describes several ways that, during the past decade, researchers in mathematics and science education have fundamentally revised their ways of thinking about: (a) the nature of mathematics or science, (b) the nature of real-life decision-making situations in which these ideas are useful, and (c) the nature of productive teaching and learning. In general, when attention focuses on students or teachers, emerging new perspectives have shifted away from machine metaphors for describing development; instead, they emphasize perspectives that are more systemic, more organic, and better aligned with constructivist philosophies about the nature of learning and development. Yet, when attention shifts beyond students and teachers to focus on programs, schools, or school systems, development continues to be described using factory-based metaphors borrowed from an industrial age.

When old-fashioned factory models are used to conceptualize the development of complex systems, both teaching and learning tend to be treated as if they were similar to the process of assembling an automobile or another inert machine. Rather than molding and shaping the activities of living systems, these systemic wholes are treated as if they were nothing more than the sum of their parts. Furthermore, it is assumed that, when someone (or something) acts on these systems, they do not act back. In particular, observers (and tests) are considered to be completely objective and detached from the systems being observed.

When machine metaphors are used to think about curricular improvements in schools, the overall process might be called a conformance model for trying to improve instruction. Such an approach begins by defining fixed conceptions of desired outcomes. Then, participants are coerced to teach to tests that tend to be treated as de facto operational definitions of these standards. But, for reasons we've described earlier in this chapter, negative characteristics of this approach arise because the scope, length, and depth of tests generally force them to restrict attention to simplistic notions of expertise and to low-level tasks that are aligned poorly with intended instructional goals.

An alternative to conformance models for curriculum change might be called planned experimentation. It is an approach that is used commonly in modern businesses; and, it tends to be especially useful in situations where: (a) rapid adaptation is needed, (b) a "one size fits all" policy is hopelessly out of date, and (c) huge centralized, top-down, management styles

are far too slow and insensitive to local needs. It also is an approach that is especially relevant to the development of nearly any community of complex adapting systems, including those that are similar to biological organisms rather than to inert machines.

As Darwin's theory of evolution suggests, diversity and variation are necessary ingredients for evolution to occur. They are not incompatible processes. In fact, if there is no diversity or no selection, then the relevant systems are not likely to evolve. Other necessary processes include selection, cross-fertilization, and preservation across time.

In the context of schools and programs of instruction, what the preceding principles imply is that the goal of assessment should be to improve performance, not merely to audit it. In particular, assessment should anticipate action (Clarke, 1996). It should be longitudinal and recursive; optimization and documentation should not be treated as if they were incompatible processes, and diversity should be encouraged as long as it is coupled with effective selection procedures to promote development.

When a planned experiment is used to improve instruction in schools, an important assumption is that, as the system develops, participants' conceptions of goals become much more clear, powerful, and sophisticated. Therefore, descriptions of expertise should not be locked in at the beginning of projects. Instead, these descriptions should be tested, revised, and refined gradually over the course of the project; and, the conceptions of excellence that evolve can be based on trails of documentation that accumulate. How can systems get better without beginning with a fixed and final definition of best? Regardless whether we are investigating complex systems that develop in education, the arts, crafts, sports, or professions like business or engineering, what we need are the abilities to elicit, observe, and assess the relevant constructs - not necessarily the ability to produce dictionary-style definitions, and certainly not the willingness to reduce goals to checklists of simple rules.

Contexts that elicit complex performances don't necessarily require the relevant participants to reveal observable ways of thinking; and, they also don't necessarily provide useful tools for comparing or assessing the quality of competing systems. Nonetheless, it often is not difficult to identify situations that require the relevant systems to be elicited and revealed, and it often is not difficult to identify ways to compare and assess the results that are produced. For example, we may not know how to define what makes Granny a great cook; however, it still may be easy to identify situations that will elicit and reveal her capabilities, and it also may be easy to compare and assess alternative results that are produced.

In general, to formulate useful operational definitions of complex, dynamic, and continually adapting systems, it is not always necessary to begin by producing a precise

dictionary-style descriptions accompanied by ways to measure distances from this ideal. Sometimes, it is sufficient to specify: (a) how to elicit the system; (b) how to observe its characteristics; and (c) how to compare the alternatives that occur. Then, when these capabilities exist, it often is possible to create a series of interactions that involve iterative, formative, feedback loops that produce a trail of documentation that reveals the nature of the evolving system.

If we apply this approach to the development of programs, schools, or school systems, the first tools that are needed are contexts that can be used to get programs (or schools, or school systems) to reveal their relevant ways of thinking. Possibilities include: (a) curricular goals that are formulated; (b) alternative assessment instruments that are developed; (c) progress reports that are created to describe achievements; (d) dissemination and implementation plans that are adopted to facilitate improvement; and (e) teacher development activities that are offered to promote change. Then, in addition to identifying the kinds of contexts that are required for eliciting and revealing relevant ways of thinking, criteria also are needed that can be used to compare or assess the products that are generated.

Fortunately, finding ways to compare or assess the effectiveness of products that are generated by schools tends to be no more mysterious than finding ways to compare and assess the results of Granny's cooking. But, in the case of schools, judgments about quality nearly always should involve students' achievement. Therefore, when carrying out such processes, some important principles to keep in mind include the following:

1. Achievement usually needs to be assessed using something more than brief tests that reduce expertise to simplistic lists of condition–action rules.

2. Students' relevant products or performances usually should include more than those that can be interpreted and assessed easily by a machine.

3. Emphasis needs to shift beyond superficial coverage of a large number of small tasks to the comprehensive treatment of a small number of big ideas.

4. Quality ratings should not ignore the conditions under which complex performances occur, and complex profiles should not be collapsed into simplistic scores on a scale that recognizes only a single dimension along which progress can be made.

## COMPLEX SYSTEMS: TO KNOW THEM IS TO CHANGE THEM

To end this chapter about ways of formulating productive operational definitions of complex constructs in mathematics and science education, it is appropriate that we find ourselves returning to issues about students' ways of thinking, and about relationships between behavioral objectives and deeper cognitive objectives of instruction. It also is appropriate that we reconsider one of the most prevalent and damaging myths related to educational assessment: the assumption that assessment does not (and should not) induce changes in the entities being assessed.

For those familiar with Heisenberg's indeterminacy principle in physics, it is a familiar notion that to measure a construct is to change it. Therefore, it should come as no surprise that similar principles also may apply to the kinds of complex, dynamic, and continually adapting systems that occur in mathematics and science education. To know them is to interact with them; and, to interact with them often involves changing them in fundamental ways. For example, in the case of studies assessing the effectiveness of teaching (or programs, or curricular materials), standardized tests often go beyond being used as neutral indicators of learning outcomes. If they are used to clarify instructional goals, they often become powerful components of the initiatives themselves (Clarke & Stephens, 1996). Consequently, far from being passive assessments of non-adapting systems, they may have powerful positive or negative effects, depending on whether they support or subvert objectives of the relevant systems.

In mathematics and science education, there is another fundamental way that assessments often change the nature of the systems being observed. That is, even at the level of students' thinking about individual problem solving situations, it tends to be undesirable (and virtually impossible) to devise assessment tasks in which no learning is likely to occur. For example, for the kind of thought-revealing activities that we have advocated in this chapter, the interpretations that students are challenged to produce inherently involve significant forms of

learning. This is because the new models and ways of thinking that develop often require previously existing conceptual systems to be integrated, differentiated, extended, or refined in significant ways. Furthermore, if the graphs, charts, diagrams, spreadsheets, and other symbolic descriptions that students generate are shareable, reusable, easily modifiable, and transportable beyond one particular situation, then they often represent significant forms of generalization and higher order thinking.

The preceding kind of thought-revealing activities emphasize that, above all, thinking mathematically involves interpreting situations by quantifying them, coordinatizing them, and mathematizing them in other ways. It is at best a half-truth to claim that elementary mathematics is about numbers and computation; it is also about quantities and quantitative relationships. Consequently, by treating mathematics as being (only) about numbers and operations with numbers, traditional textbooks and tests have tended to trivialize the interpretation stages of problem solving. For example, the results that students are requested to produce seldom include descriptions, explanations, or justifications in which they must reveal and test explicitly how they interpret problem solving situations; and, even in cases where a mathematical interpretation needs to be generated, the kinds of quantification that are involved rarely go beyond simple counts and measures. Yet, the kinds of situations that children should be able to use numbers to describe include those involving ranks and positions (coordinates), signed quantities (such as those involving negatives and positives), quantities assigned to qualitative information (such as those involving weights or values), multivalued quantities (such as those involving magnitude and direction), relationships among quantities (ratios), actions or transformations (such as those involving stretching, shrinking, or exchanging operations), changing quantities (rates involving "per" quantities), intensive quantities (such as those involving "-ness" quantities like denseness, and speediness), or accumulating quantities.

## SUMMARY AND CONCLUSIONS

An abiding goal of education is the specification of the desired outcomes of instruction in observable and assessable forms. As our conception of valued outcomes becomes more sophisticated, as we move from behavioral objectives to cognitive objectives, as we come to acknowledge the complexity of the settings in which we work, and as we come to recognize

the symbiotic character of the constructs that we seek to assess within their encompassing systems, so our assessment (and research) must develop in subtlety and complexity.

In this chapter, analogies with detection systems from physical science have been used to point to similarities in the challenges posed by the study of constructs only observable indirectly, whether in educational systems or in physical systems. These similarities are important. They allow us to benefit from conceptual advances made in other disciplines and to validate the rigor of our own methods through their correspondence with the endorsed practices of other communities.

The identification of the three-fold character of model-eliciting activities provides a generalizable structure for the practical manifestation of valued educational constructs. This structure focuses attention (as classroom practitioners or as educational researchers, for instance) on activities that create situations conducive to the constructs to be studied, requires products or observation tools through which the construct is made manifest, and produces these products in a form amenable to assessment against specified criteria for quality. In their combination, these three elements—situations, products, and quality assessment criteria—constitute the operationalization of the construct in question.

This chapter has identified cognitive objectives as appropriate goals for contemporary instruction. Ways of thinking represent an elaboration of this conception of cognitive objectives in a form applicable to individuals and to educational systems. Multitiered teaching experiments represent one approach to the study of such systems; they are discussed in greater detail in other chapters.

Starting from an acknowledgment of the inevitable perturbation of the observed system in the act of assessment, perhaps the most contentious of the suggestions put forward in this chapter is advocacy of the use of assessment to change the observed system purposefully toward the desired outcomes. In this, the authors anticipate the development of assessment systems that both reveal valued constructs and deliberately prompt adaptations within the systems studied in directions advantageous to the system or to the individual whose performances are assessed.

## APPENDIX A: Quality Assurance Guide

This Quality Assurance Guide is designed to help teachers (and students) evaluate the products that are developed in response to model-eliciting activities with the following characteristics: the goal is to develop a conceptual tool, (b) the client who needs the tool is clearly identified, (c) the client's purposes are known, and (d) the tool must be sharable with other people and must be useful in situations where the data are different than those specified in the problem.

| ***Performance Level*** | ***How useful is the product?*** | ***What might the client say?*** | ***What questions should be asked?*** |
|---|---|---|---|
| Requires Redirection | The product is on the wrong track. Working longer or harder won't work. The students may require some additional feedback from the teacher. | *"Start over. This won't work. Think about it differently. Use different ideas or procedures."* | To assess students' work, put yourself in the role of the client. To do this, it's necessary to be clear about answers to the following questions.<br>1. Who is the client?<br>2. What conceptual tool does the client need?<br>3. What does the client need to be able to do with the tool?<br><br>Then, the quality of students' work can be determined by focusing on the question—*How useful is the tool for the purposes of the client?*<br><br>To assess usefulness, and to identify strengths and weaknesses of different results that students produce, it would be helpful to consider the following questions.<br>1. What information, relationships, and patterns does the tool take into account?<br>2. Were appropriate ideas and procedures chosen for dealing with this information?<br>3. Were any technical errors made in using the preceding ideas and procedures?<br><br>But, the central question is—*Does the product meet the client's needs?* |
| Requires Major Extensions or Refinements | The product is a good start toward meeting the client's needs, but a lot more work is needed to respond to all of the issues. | *"You're on the right track, but this still needs a lot more work before it'll be in a form that's useful."* | |
| Requires Only Minor Editing | The product is nearly ready to be used. It still needs a few small modifications, additions, or refinements. | *"Hmmm, this is close to what I need. You just need to add or change a few small things."* | |
| Useful for this Specific Data Given | No changes will be needed to meets the immediate needs of the client. | *"Ahhh, this will work well as it is. I won't even need to do any editing."* | |
| Sharable or Reusable | The tool not only works for the immediate situation, but it also would be easy for others to modify and use it in similar situations. | *"Excellent, this tool will be easy for me to modify or use in other similar situations—when the data are slightly different."* | The product should make it clear that: (1) The students went beyond producing a tool that *they* themselves could use to also produce a tool that *others* could use—by including needed explanations, and by making it as simple, clear, and well organized as possible.; (2) the students went beyond thinking *with* the tool to also think *about* it—by identifying underlying assumptions (so that others know when the tool might need to be modified for use in similar situations); (3) the students went beyond *blind* thinking to also think *about* their thinking—by recognizing strength and weaknesses of their approach compared with other possible alternatives. |



BEST COPY AVAILABLE

This guide is the fourth of four tools that are useful for making sense of students' work related to model-eliciting activities. The first tool is an "Observation Sheet" that can be used to observe the processes that students' thinking goes through to develop final responses. The second tool is a "Ways of Thinking Sheet" that can be used to sort out strengths and weaknesses of the alternative results that students produce. The third tool is a "Self Reflection Sheet" that helps students reflect about roles they may have played during problem solution, feelings, and attitudes that may have influence performance.

## APPENDIX B

### The Softball Problem

To begin the Softball Problem, teachers often read the newspaper article and discuss the warm-up questions on the day before they want students to begin work on the main activity. This helps students become familiar with the situation before they are expected to begin work, and it also eliminates a lot of "wheel spinning" during the actual problem solving session.

### The 'Crack' of the Bat Rings Throughout Central Indiana Town and Across Fields During Summer Evenings

Summer is underway which means only one thing to the people in town of Gasburg: softball season has arrived. For many children in this Central Indiana town, softball is an annual tradition where players learn about competition, friendship and establish many fond memories.

But some parents are trying to make changes to the Gasburg Town Summer League, which has basically remained unchanged since the league was founded in 1946. The main change which parents want is equal competition between the teams. And they want it now.

This season—like seasons past—one team will most likely run away with the league title. Each year, one coach seems to get the best players in the town to be on the same team, which demolishes the competition. "Last year the Rib Shack won all of their games. Their closest win was a 10-3 win," said Kelsey Emonds. Emonds, a four-year veteran outfielder in the league, will compete in her final season before moving on to play softball in high school next year. "It's kind of sad. One part of me wants to play at this park forever and another part of me wants to move on to high school softball this season because I am tired of the unfair teams."

During her four years in the league, Emonds was on a team that breezed through the competition and then re-assigned the next year to a team that "was the joke of the league," said Emonds.

"I remember for the first three years, I was on a team with the neighborhood girls and we all played softball competitively," said Emonds. "Then I was placed on this new team which had players that couldn't play very well. It was a really bad year."

The team only won one game all season long in the 25-game season. "I just hope we have a better year. It's my last year and I want to go out with a better record and have a chance at winning a championship."

Mike Browning, league president, has expressed a concern about how teams are chosen and is trying to establish a new method of dividing interested players into fair teams.

"Attendance has dropped a lot," said Browning. "It's really not asking much to want to see a game in which the two teams are actually competing against one another."

Some changes might be coming in the way teams are assigned. Every spring, the league holds tryouts and practice games before the season starts. At the tryouts, players are tested in three different events: batting, throwing, and base running.

Currently, each player will get 10 at-bats at the tryouts. Players will be timed running from first base to second base three times and will have five chances to throw from centerfield after catching a ball.

If Browning and other parents have their way, the current system of having coaches pick the players would be abolished and the league would assign the players every year.

"The softball league hasn't changed much in over 50 years," said Browning. "If we don't make changes now, people will be so turned off to softball that we won't exist in another 10 years."

## Warm-up Questions

Answer the following questions after reviewing the article and the data.

1. Why are people, e.g. parents, coaches and players, upset with the current system?
2. Who is Mike Browning?
3. Who is Kelsey Emonds?
4. What team dominated the league last year?
5. Has this been the case in other instances in the past?
6. Who threw the farthest of all the participants?

7. Who ran the fastest of all the participants?
8. Did the same person throw the farthest and run the fastest also?
9. Who had the highest batting average?

TABLE B1

| *Person* | *Throw in ft.* | | *Batting* | | | | *Run in seconds* |
|---|---|---|---|---|---|---|---|
| Tisha | 81 | 65 | double | Single | single | pop-fly | 7.2 |
| | 79 | 61 | single | pop fly | pop fly | single | 6.9 |
| | 69 | | pop-fly | sac fly | | | 6.9 |
| Kim | 67 | 61 | Pop-fly | Pop-fly | Sac-fly | Home | 5.1 |
| | 66 | 59 | Sac-fly | Sac-fly | Triple | run | 5.3 |
| | 66 | | Triple | Ground out | | Double | 5.1 |
| Sally | 83 | 75 | Home run | Ground out | Pop-fly | Sac-fly | 5.9 |
| | 80 | 82 | Single | Double | Ground-out | Double | 6. |
| | 80 | | Sac-fly | Pop-fly | | | 6.2 |
| Barb | 97 | 89 | Triple | Ground-out | Pop-fly | Sac-fly | 6.4 |
| | 98 | 90 | Ground-out | Ground-out | Single | Sac-fly | 6.3 |
| | 91 | | Single | Pop-fly | | | 6.3 |
| Melissa | 98 | 88 | Pop-fly | Home run | Single | Double | 6.3 |
| | 90 | 86 | Double | Ground-out | Sac-fly | Ground- | 6.5 |
| | 84 | | Sac-fly | Pop-fly | | out | 6.7 |
| Charlene | 86 | 77 | Single | Strike out | Pop-fly | Ground- | 6.5 |
| | 83 | 80 | Double | Ground out | Pop-fly | out | 7.1 |
| | 65 | | Strike-out | Sac-fly | | Sac-fly | 6.5 |
| Cynthia | 72 | 72 | Strike out | Pop-fly | Strike-out | Double | 6.9 |
| | 72 | 69 | Ground-out | Strike-out | Double | Strike- | 6.1 |
| | 71 | | Strike-out | Pop-fly | | out | 6.4 |
| Rebecca | 49 | 56 | Strike-out | Strike-out | Single | Strike- | 5.5 |
| | 52 | 54 | Strike-out | Pop-fly | Double | out | 5.7 |
| | 56 | | Ground-out | Double | | Single | 5.6 |
| Beth | 61 | 54 | Single | Sac-fly | Ground-out | Pop-fly | 5.9 |
| | 61 | 59 | Double | Triple | Single | Home- | 6.0 |
| | 57 | | Strike-out | Ground-out | | run | 5.9 |
| Mary | 53 | 50 | Strike-out | Strike-out | Pop-fly | Triple | 6.5 |
| Pat | 57 | 37 | Sac-fly | double | Single | Single | 6.2 |
| | 52 | | Sac-fly | sac-fly | | | 6.4 |
| Kathy | 98 | 90 | Strike-out | Pop-fly | Double | Strike- | 5.6 |
| | 97 | 92 | Single | Pop-fly | Pop-fly | out | 5.9 |
| | 99 | | single | Sac-fly | | Sac-fly | 5.4 |
| Esther | 90 | 87 | Strike-out | Strike-out | Single | Sac-fly | 6.1 |
| | 90 | 85 | Strike-out | Double | Single | Double | 6.4 |
| | 86 | | Sac-fly | Sac-fly | | | 6.2 |
| Kelly | 80 | 80 | Home-run | Strike-out | Home-run | Pop-fly | 5.2 |
| | 79 | 82 | Strike-out | Double | Strike-out | Pop-fly | 5.1 |
| | 77 | | Strike-out | Triple | | | 4.9 |

Key For Hitting Evaluation

1. A strike-out occurs when a person gets three strikes before she hits the ball or gets four balls (a walk).
2. A pop-fly occurs when a person hits the ball in the air and another individual catches the ball before it hits the ground. An out is achieved.
3. A ground-out occurs when a person hits the ball and a fielder catches it and throws it to first before the batter arrives. An out is achieved.
4. A Sac-fly occurs when a person hits the ball in the air and a person catches it. Unlike a pop-fly a base-runner advances to the next base.
5. A single occurs when a person hits the ball and she gets to first base before the ball arrives.
6. A double occurs when a person hits the ball and she gets to second base before the ball arrives.
7. A triple occurs when a person hits the ball and she gets to third base before the ball arrives.
8. A home-run occurs when a person hits the ball over the fence or when a person hits the ball and arrives home before the ball arrives.

Coaches' Comments

Tisha: She has great fielding skills. Also she is very easy to work with. Her father was a professional baseball player.

Kim: Very fast athlete. She can get around the bases quickly and to balls quickly. She has gotten along well with teammates in the past.

Sally: No comments available.

Barb: She has a very strong arm and she is capable of playing several positions in a game.

Melissa: Fields exceptionally well. Her batting was improved near the end of the year.

Charlene: She's very mature for her age. When her coordination catches up with her age, she could be quite good.

Cynthia: Does many things well.

Rebecca: She knows more about the game than anyone else that I've coached. Her father played college baseball.

Beth: She's a very good offensive weapon and she plays well in the outfield.

Mary Pat: She does many things well, but she doesn't stand out in one area.

Kathy: She has a great throwing arm. This was the first year she ever played and she has the potential to become a great athlete.

Esther: No comments available.

Kelly: She helps the younger athletes. However, she has been known to miss games.

**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*

# NOTICE

## Reproduction Basis

| | |
|---|---|
| X | This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form. |
| | This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket"). |